Coalescent Histories for Caterpillar-Like Families

Noah A. Rosenberg

Abstract—A coalescent history is an assignment of branches of a gene tree to branches of a species tree on which coalescences in the gene tree occur. The number of coalescent histories for a pair consisting of a labeled gene tree topology and a labeled species tree topology is important in gene tree probability computations, and more generally, in studying evolutionary possibilities for gene trees on species trees. Defining the *T_r-caterpillar-like family* as a sequence of n-taxon trees constructed by replacing the r-taxon subtree of n-taxon caterpillars by a specific r-taxon labeled topology T_r, we examine the number of coalescent histories for caterpillar-like families with matching gene tree and species tree labeled topologies. For each T_r with size $r \leq 8$, we compute the number of coalescent histories for n-taxon trees in the T_r-caterpillar-like family. Next, as $n \to \infty$, we find that the limiting ratio of the numbers of coalescent histories for the T_r family and caterpillars themselves is correlated with the number of labeled histories for T_r. The results support a view that large numbers of coalescent histories occur when a tree has both a relatively balanced subtree and a high tree depth, contributing to deeper understanding of the combinatorics of gene trees and species trees.

Index Terms—Combinatorial identities, labeled histories, labeled topologies, lineage sorting, phylogenetics

1 Introduction

A *coalescent history* is a list of edges of a species tree topology on which the coalescences in a gene tree topology take place. A pair consisting of a labeled gene tree topology G and a labeled species tree topology S, both with n leaves, specifies a set of possible coalescent histories, each of which gives a distinct pairing of coalescences in G with edges of S (see Fig. 1). Each pairing must satisfy a series of rules that constrain the evolution of gene trees conditional on species trees.

Coalescent histories arise in the study of the combinatorics of gene trees and species trees. A key result is that under the "multispecies coalescent" [1], a standard probability model for genealogical evolution, the probability conditional on a species tree σ with labeled topology S and branch lengths λ that the labeled topology G of a random gene tree is g can be written as

$$\mathbb{P}_\sigma[G = g] = \sum_{h \in H(G,S)} \mathbb{P}_\sigma[G = g, h], \tag{1}$$

where $H(G,S)$ denotes the set of coalescent histories for the pair (G,S) [2]. The use of coalescent histories separates the evaluation of the marginal probability $\mathbb{P}_\sigma[G = g]$ into two problems: the simpler computation of the joint probability $\mathbb{P}_\sigma[G = g, h]$, and the enumeration of the coalescent histories in $H(G,S)$.

As a central component of the mathematical relationship between gene trees and species trees, coalescent histories have been important in a variety of contexts. Owing to the structure of (1), the number of coalescent histories influences the computational complexity of the evaluation of the probability of a labeled gene tree topology [2], [3], [4]. The collection of joint probabilities $\mathbb{P}_\sigma[G = g, h]$, obtained using an enumeration of coalescent histories, assists in numerical characterizations of features of gene tree probability distributions [5]. Coalescent histories have been employed as part of proofs for properties of algorithms that infer species trees from gene trees [6], [7]. They can describe a state space for models that examine changes in gene tree topologies along a genome [8], [9], [10]. Finally, they appear in studies of hybridization detection, where they are defined in relation to species networks rather than trees [11].

In introducing coalescent histories, Degnan and Salter [2] obtained the set $H(G,S)$ by exhaustively considering elements of a larger superset that encoded a collection of constraints weaker than those that characterize gene tree evolution. Rosenberg [3] and Than et al. [4] then reported faster enumeration algorithms by precisely identifying the set of coalescent histories without requiring examination of the larger superset. These studies have shown that the number of coalescent histories associated with a labeled gene tree topology G and labeled species tree topology S can be counted using a recursive formula [3]. Investigation of this recursion then provides a basis for analysis of the mathematical properties of the number of coalescent histories [3], [4], [12].

Of particular interest are families of pairs (G,S) for which the recursion is solvable exactly, so that the number of coalescent histories can be studied nonrecursively. When $G = S$ and both G and S have an n-taxon *caterpillar* shape—a topology that possesses one internal node descended from all other internal nodes (see Fig. 2A)—Degnan [13] demonstrated that the number of coalescent histories is the Catalan number $C_{n-1} = \binom{2n-2}{n-1}/n$. Rosenberg [3] further showed that the corresponding number of coalescent histories for an n-taxon *pseudocaterpillar*—a topology with

• *The author is with the Department of Biology, Stanford University, Stanford, CA. E-mail: noahr@stanford.edu.*

Manuscript received 26 June 2013; revised 16 Sept. 2013; accepted 18 Sept. 2013; published online 30 Sept. 2013.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2013-06-0189. Digital Object Identifier no. 10.1109/TCBB.2013.123.

1545-5963/13/$31.00 © 2013 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

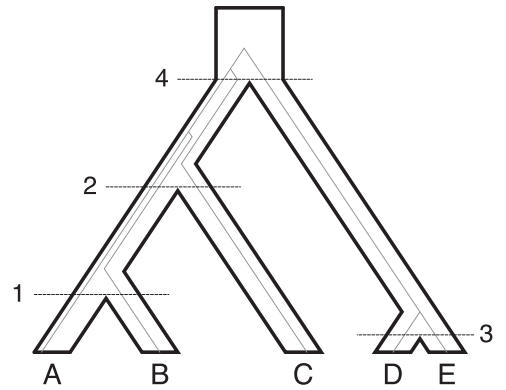

Fig. 1. A coalescent history for a five-taxon labeled gene tree topology and labeled species tree topology. Internal nodes of the species tree are numbered according to a postorder traversal, and the edge above a node is identified with the node. The species tree, represented by thick lines, and the gene tree, represented by thin lines, both have labeled topology (((A,B),C),(D,E)). Coalescences (A,B), ((A,B),C), (D,E), and (((A,B),C),(D,E)) occur on edges 2, 4, 3, and 4, respectively. The coalescent history shown is one of 10 possible for the given gene tree and species tree labeled topologies.

a four-taxon symmetric subtree whose root descends from all internal nodes that are not part of the subtree (see Fig. 2B)—is $(5n-12)C_{n-1}/(4n-6)$. As $n \to \infty$, the ratio of the numbers of coalescent histories for n-taxon pseudocaterpillars and caterpillars approaches $5/4$.

This result illustrates a principle that a large number of coalescent histories can be produced when multiple sequences of coalescences are permitted, and when many branches exist on which those coalescences can occur [3], [12]. A pseudocaterpillar has two possible sequences of coalescences—either one pair of lineages forming a cherry coalesces first, or the other does—whereas a caterpillar has only one possible coalescence sequence. However, the depth of an n-taxon caterpillar—the greatest distance from a leaf to the root—exceeds that of an n-taxon pseudocaterpillar by 1. As n becomes large, the increase in the number of coalescent histories owing to the extra coalescence sequence for a pseudocaterpillar more than compensates for the decrease in coalescent histories caused by its smaller depth compared to a caterpillar, and the pseudocaterpillar has more coalescent histories.

For a fixed n, this example highlights conflicting trends. Larger numbers of coalescence sequences occur for certain relatively balanced tree topologies described as *maximally probable* by Degnan and Rosenberg [14]. At the same time, balanced trees are not as deep as caterpillars and have fewer branches on which a typical coalescence can occur. As n increases, does the larger number of coalescence sequences for such trees overcome the reduction in depth, as it does for pseudocaterpillars?

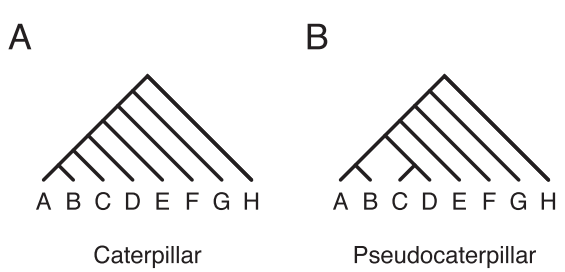

Fig. 2. Caterpillar and pseudocaterpillar labeled topologies with $n = 8$ leaves. (A) A caterpillar tree. (B) A pseudocaterpillar tree.

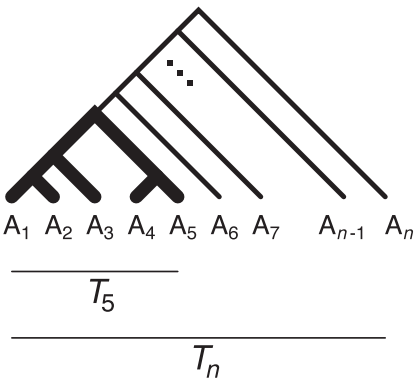

Fig. 3. T_r-caterpillar-like families. T_5 represents a choice of labeled topology T_r with $r = 5$ leaves, and T_n represents a caterpillar tree with its five-taxon subtree replaced by T_5.

Here, this question is posed for small caterpillar-like families, with $G = S$. A *T_r-caterpillar-like family* of trees based on labeled r-taxon subtree T_r consists of a sequential set of labeled trees of size $n \geq r$ in which the r-taxon caterpillar subtree of an n-taxon caterpillar is substituted by T_r (see Fig. 3). The two simplest caterpillar-like families are the family of caterpillars themselves and the pseudocaterpillars, in which the r-taxon subtree is a four-taxon symmetric subtree. All caterpillar-like families in which the specified subtree T_r has size $r \leq 8$ are considered, and for each family, both a general n-taxon formula for the number of coalescent histories and the $n \to \infty$ limit of the number of coalescent histories in relation to the Catalan number C_{n-1}, the number of coalescent histories for the n-taxon caterpillar, are obtained.

2 Preliminaries

The notation here generally follows Rosenberg [3] and Rosenberg and Degnan [12]. To formally define a coalescent history, following Than et al. [4] and Rosenberg and Degnan [12], consider a binary rooted tree topology T with n leaves labeled by set X, and with internal edges $E(T)$. Numbers are assigned to the nodes and edges of T, identifying each node with its immediate ancestral edge (see Fig. 1). These assignments are ordered according to a postorder traversal, so that the number for a descendant edge is smaller than the numbers for all its ancestral edges. Define a partial order $\leq_T$, by which two distinct edges e_1 and e_2 satisfy $e_1 <_T e_2$ if and only if e_2 is ancestral to e_1 in T. For each internal edge e in $E(T)$, including the edge ancestral to the root, let c_e^T denote the *cluster* of T identified with edge e. This cluster consists of the set of labels in X for all leaves descended from e. The set of clusters in T, $\{c_e^T : e \in E(T)\}$, is denoted C_T. Any pair of distinct clusters is either disjoint or nested, such that one is a supercluster and the other is a subcluster.

Definition 1. *For a labeled gene tree topology G and a labeled species tree topology S with the same set of leaf labels, a* coalescent history *is a mapping $\alpha : C_G \to E(S)$ such that 1) for each $Y \in C_G$, $Y \subset c^S_{\alpha(Y)}$, and 2) for each $e_1, e_2 \in E(G)$, if $e_1 <_G e_2$, then $\alpha(e_1) \leq_S \alpha(e_2)$.*

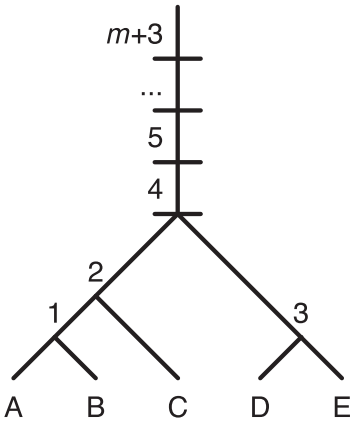

Fig. 4. An m-extended species tree topology for which the edge above the root is artificially divided into m edges. The numbers denote labels for the edges. If the gene tree topology is (((A,B),C),(D,E)), then an m-extended coalescent history involves a coalescence of ((A,B),C) and (D,E) on an edge k in $\{4,\ldots,m+3\}$, a coalescence of D and E on an edge in $\{3,\ldots,k\}$, a coalescence of (A,B) and C on an edge ℓ_1 in $\{2,4,\ldots,k\}$, and a coalescence of A and B on an edge ℓ_2 in $\{1,2,4,\ldots,k\}$ satisfying $\ell_2 \leq \ell_1$. If $m=1$, then the edge above the root is not subdivided, and m-extended coalescent histories are equivalent to coalescent histories.

The first condition specifies that in a coalescent history, each cluster Y of the gene tree topology coalesces at the most recent common ancestor (MRCA) of Y in the species tree topology, or deeper than this MRCA. The second condition specifies that cluster Y cannot find its MRCA on an edge deeper than an edge on which one of its superclusters finds its MRCA.

An *m-extended coalescent history* for a labeled gene tree topology and labeled species tree topology is a coalescent history for the gene tree topology and species tree topology when the edge above the root of the species tree topology is subdivided into m components (see Fig. 4). Denote the number of m-extended coalescent histories for a gene tree topology G and a species tree topology S, when $G=S$, by $A_{S,m}$. According to [3, Theorem 3.1], $A_{S,m}$ can be obtained by a recursion:

$$A_{S,m} = \sum_{k=2}^{m+1} A_{S_L,k} A_{S_R,k}, \tag{2}$$

where for all $m \geq 1$, $A_{S,m}=1$ if S has only one taxon. In the recursion, S_L and S_R represent the "left" and "right" subtrees of S. By convention, choose S_L and S_R such that the number of leaves of S_L is greater than or equal to that of S_R. The number of coalescent histories when the gene tree and species tree both have labeled topology S, or $A_{S,1}$, is obtained as the $m=1$ case of the number of m-extended coalescent histories.

If S is part of the same caterpillar-like family as S_L, then S_R has one taxon, and $A_{S_R,k}=1$ for all k. Consequently, for successive trees in the same caterpillar-like family, (2) simplifies. Consider a T_r-caterpillar-like family, where T_r is a labeled topology with r taxa. Denote the members of this family by $\{T_n\}_{n\geq r}$. For each $n>r$, the left subtree of T_n is T_{n-1}, and the right subtree of T_n has only one taxon. By (2),

$$A_{T_n,m} = \sum_{k=2}^{m+1} A_{T_{n-1},k}. \tag{3}$$

The base case of this recursion is a condition that specifies the values of $A_{T_r,m}$ for all m. By iterating (2), Rosenberg [3] obtained formulas for $A_{S,m}$ for all labeled topologies S with $n \leq 9$ taxa. Thus, for each T_r with $r \leq 9$, the base case $A_{T_r,m}$ is reported in [3, Tables 1, 2, 3, and 4].

For the two simplest T_r-caterpillar-like families—caterpillars and pseudocaterpillars—Rosenberg [3] used (3) to compute the number of m-extended coalescent histories for n-taxon T_r-caterpillar-like trees. For caterpillars, $r=2$, and for $n \geq 2$ and $m \geq 1$ [3, Theorem 3.4],

$$A_{T_n,m} = \frac{m}{(n-1)!}\frac{(m+2n-3)!}{(m+n-1)!}. \tag{4}$$

In the pseudocaterpillar case, $r=4$, and for $n \geq 5$ and $m \geq 1$ [3, Theorem 3.7],

$$A_{T_n,m} = \frac{m}{(n-1)!}\frac{(m+2n-5)!}{(m+n-1)!}[2m^2 + (5n-11)m + (5n^2-22n+21)]. \tag{5}$$

As in [3] and [12], without loss of generality, a single labeling is taken here as representative of each unlabeled species tree topology. Thus, it is possible to study *the* caterpillar and pseudocaterpillar families by considering an arbitrary labeling in each family, with each successive taxon in the family providing an additional taxon label. When the labeling is not needed, the arbitrarily labeled species tree topology is abbreviated by its unlabeled shape, and the labeled and unlabeled topologies are considered interchangeably.

We examine properties of the number of coalescent histories in a T_r-caterpillar-like family in relation to two aspects of T_r. The first is its *rank*, which for a given unlabeled binary tree shape is its position in an enumeration of all unlabeled binary tree shapes with r taxa. In the enumerated list of shapes in canonical form, for each internal node, at least as many taxa appear in the left subtree of the root as in the right subtree. Shapes with more taxa in their left subtrees have lower rank than do shapes with fewer taxa in their left subtrees. For two shapes with equally many taxa in their left subtrees, the rank is smaller for the shape whose left subtree has a lower rank; if the left subtrees are identical, then the rank is smaller for the shape whose right subtree has a lower rank. In each shape in canonical form, when the two subtrees of a node have equally many leaves but are not identical, the left subtree has lower rank. The caterpillar shape has rank 1, and balanced shapes tend to have high rank. This ranking scheme is equivalent to that of Furnas [15, Section 2.5.1.1], except that its canonical form places more taxa in the left rather than the right subtree.

The second variable is the number of *coalescence sequences*, or *labeled histories*, for labeled topology T_r, where in distinct labeled histories, the coalescences in the labeled topology are identical, but when the topology is viewed as having been generated temporally, from the leaves toward the root, the coalescences occur in a different order. The number of labeled histories for T_r is

$$N(T_r) = \frac{(r-1)!}{\prod_{j=3}^{r}(j-1)^{d_j(T_r)}}, \tag{6}$$

where $d_j(T_r)$ is the number of internal nodes of T_r from which exactly j leaves descend [16], [17]. It can now be stated that T_r is an *r-maximally probable* labeled topology if $N(T_r) \geq N(T_r')$ for all labeled topologies T_r' that also have r taxa. Letting $z = 1 + \lfloor \log_2[(r-1)/3) \rfloor \rfloor$, the r-maximally probable topologies can be characterized recursively as those topologies whose two subtrees immediately descended from the root are 2^z- and $(r-2^z)$-maximally probable [18], [19], [20]. For fixed r, a caterpillar T_r has the fewest coalescence sequences, $N(T_r) = 1$, and labeled topologies T_r with balanced shapes tend to have high $N(T_r)$.

3 Results

Generalizing the approach used previously for obtaining (4) and (5), we establish a procedure for obtaining the number of m-extended coalescent histories for T_r-caterpillar-like trees with a given choice of T_r (see Section 3.1). We next illustrate this procedure with a specific T_r (see Section 3.2). By applying the procedure, we compute the number of m-extended coalescent histories for all choices of T_r with $r \leq 8$ (see Section 3.3). For each T_r, by setting $m = 1$, the number of coalescent histories is obtained, and the ratio of this quantity to the number of coalescent histories for an n-taxon caterpillar is used to relate the number of coalescent histories for the T_r-caterpillar-like family to that of the caterpillar family.

3.1 The General Procedure

Consider some T_r with $r \geq 2$ taxa. If $r = 2$, then T_r is the first member of the caterpillar family (it is convenient to begin at $r = 2$ rather than at the trivial case of $r = 1$). Otherwise, if the right subtree of T_r has exactly one taxon, then T_r and T_{r-1} are in the same family, and the T_r-caterpillar-like family is a subset of the T_{r-1}-caterpillar-like family. We can then consider the T_{r-1}-caterpillar-like family in place of the family that starts with T_r. Once the minimal element of a T_r-caterpillar-like family is identified, we perform the following series of steps.

1. $A_{T_r,m}$ is obtained by iteration of (2). For $r \leq 9$, $A_{T_r,m}$ has already been reported in [3, Tables 1, 2, 3, and 4].
2. Equation (3) is iterated approximately $2r$ times to obtain formulas for $A_{T_{r+1},m}, A_{T_{r+2},m}, \ldots$. The number of iterations is chosen to be sufficient to identify a general pattern in the formulas.
3. Using the formulas obtained in Step 2, a pattern is suggested for the general formula for $A_{T_n,m}$, where $n \geq n_0$ for some $n_0 \geq r$.
4. Induction is used to prove that the pattern in Step 3 is correct.
5. In the formula for $A_{T_n,m}$, m is set to 1 to obtain the number of coalescent histories $A_{T_n,1}$.
6. To obtain the limiting ratio of the numbers of coalescent histories for the T_r-caterpillar-like family and the caterpillar family itself, $\lim_{n\to\infty}(A_{T_n,1}/C_{n-1})$ is computed.

Two of the steps are nontrivial. First, in Step 3, a pattern must be suggested for $A_{T_n,m}$. For each of the choices of T_r that I have considered (all T_r with $r \leq 8$), I have found by application of Step 2 that there exists $n_0 \geq r$ such that for all $n \geq n_0$, the formula for $A_{T_n,m}$ can be written

$$A_{T_n,m} = \frac{m}{(n-1)!}\frac{(m+2n-c)!}{(m+n-1)!}\sum_{i=0}^{c-3} a_{c-3-i}(n)m^i. \tag{7}$$

Here, $c \leq n_0 + 1$ is an integer and $a_i(n)$ represents a polynomial of degree i in n. We will see for a particular T_r in the next section how such a formula can be proposed, with c and the $a_i(n)$ specified. Note that because a polynomial $a_i(n)$ of degree $c-3-i$ is uniquely determined by $c-2-i$ points $(n_1, a_i(n_1)), (n_2, a_i(n_2)), \ldots, (n_{c-2-i}, a_i(n_{c-2-i}))$, to propose a general pattern for a_{c-3-i}, we require $A_{T_n,m}$ to be computed for each n from n_0 to $n_0 + c - 3 - i$. For the polynomial that requires the largest n for proposing this pattern, namely the $i = 0$ case, the largest n required is $n_0 + c - 3$. For each T_r with $r \leq 8$, I have observed that $n_0 + c - 3 \leq 3r$, so that in all cases I have examined, at most $2r$ iterations are sufficient in Step 2 to propose the general formula in Step 3—and often, the number of iterations required is closer to r than to $2r$.

In Step 4, the proposed formula for $A_{T_n,m}$ must be verified by induction. By construction, c and the $a_i(n)$ are chosen such that the proposal is correct in the base case of $n = n_0$. It only remains to show that if the formula is correct for a given n, then it is also correct for $n+1$. Using (3), we must show that for $A_{T_n,m}$ according to (7), $A_{T_{n+1},m} = \sum_{k=2}^{m+1} A_{T_n,k}$. We now demonstrate that the sum $\sum_{k=2}^{m+1} A_{T_n,k}$ can be simplified into a closed form, reducing the problem to verifying the algebraic equivalence of $A_{T_{n+1},m}$ according to (7) and the closed form for the sum.

We begin from (7) and denote by $a_{i,j}$ the coefficient of n^j in $a_i(n)$:

$$\sum_{k=2}^{m+1} A_{T_n,k} = \sum_{k=2}^{m+1}\frac{(k+2n-c)!}{(n-1)!(k+n-1)!} \times \sum_{i=0}^{c-3}\sum_{j=0}^{c-3-i} a_{c-3-i,j}k^{i+1}n^j. \tag{8}$$

Recalling that $c \leq n_0 + 1 \leq n + 1$, so that $n - c + 1 \geq 0$, we multiply the right-hand side by $(n-c+1)!/(n-c+1)!$ to obtain

$$\sum_{k=2}^{m+1} A_{T_n,k} = \frac{(n-c+1)!}{(n-1)!}\sum_{k=2}^{m+1}\binom{k+2n-c}{n-c+1} \times \sum_{i=0}^{c-3}\sum_{j=0}^{c-3-i} a_{c-3-i,j}k^{i+1}n^j. \tag{9}$$

By repeated application of polynomial long division, $\sum_{i=0}^{c-3}\sum_{j=0}^{c-3-i} a_{c-3-i,j}k^{i+1}n^j$, treated as a polynomial of degree $c-2$ in k, can be written in terms of $(k+2n-c+1)(k+2n-c+2)\cdots(k+2n-2)$, $(k+2n-c+1)$ $(k+2n-c+2)\cdots(k+2n-3), \ldots,$ $(k+2n-c+1)(k+2n-c+2)$, $k+2n-c+1$. In other words, we can write

TABLE 1
Number of m-Extended Coalescent Histories for the T_5-Caterpillar-Like Family, Where T_5 Is a Five-Taxon Tree with Three Taxa on One Side of the Root and Two on the Other

Number of taxa n	Number of m-extended coalescent histories $A_{T_n,m}$	Number of coalescent histories $A_{T_n,1}$
6	$\frac{1}{120}m(3m^4+60m^3+445m^2+1560m+2372)$	37
7	$\frac{1}{720}m(m+7)(3m^4+78m^3+759m^2+3552m+7308)$	130
8	$\frac{1}{5040}m(m+8)(m+9)(3m^4+96m^3+1155m^2+6738m+17376)$	453
9	$\frac{1}{40320}m(m+9)(m+10)(m+11)(3m^4+114m^3+1633m^2+11394m+35240)$	1584
10	$\frac{1}{362880}m(m+10)(m+11)(m+12)(m+13)(3m^4+132m^3+2193m^2+17796m+64116)$	5577
n	$\frac{m}{(n-1)!}\frac{(m+2n-7)!}{(m+n-1)!}[3m^4+(18n-48)m^3+(41n^2-219n+283)m^2+(46n^3-369n^2+947n-774)m+(23n^4-246n^3+947n^2-1548n+896)]$	$\frac{23n^2-131n+180}{4(2n-3)(2n-5)}C_{n-1}$

$$\sum_{i=0}^{c-3}\sum_{j=0}^{c-3-i} a_{c-3-i,j}k^{i+1}n^j = \sum_{i=0}^{c-2}\frac{(k+2n-c+i)!}{(k+2n-c)!}b_i(n), \quad (10)$$

where $b_i(n)$ is a polynomial in n of degree at most $c-2$ that is computed through the long division process. Inserting (10) into (9), we then have

$$\sum_{k=2}^{m+1} A_{T_n,k} = \sum_{i=0}^{c-2} g_i(n)\sum_{k=2}^{m+1}\binom{k+2n-c+i}{n-c+i+1}, \quad (11)$$

where $g_i(n) = (n-c+i+1)!\, b_i(n)/(n-1)!$ is a rational function of n. The inner sum can be evaluated by a standard identity, reproduced as [3, Lemma 3.6], by which for $m, n_1, n_2 \geq 0$,

$$\sum_{k=0}^{m}\binom{k+n_1}{n_2} = \binom{m+n_1+1}{n_2+1} - \binom{n_1}{n_2+1}. \quad (12)$$

Using this identity,

$$\sum_{k=0}^{m+1}\binom{k+2n-c+i}{n-c+i+1} = \binom{m+2n-c+i+2}{n-c+i+2} - \binom{2n-c+i}{n-c+i+2}. \quad (13)$$

Subtracting the $k=0$ and $k=1$ terms from both sides and applying the identity $\binom{n}{k}+\binom{n}{k+1}=\binom{n+1}{k+1}$ twice, we obtain

$$\sum_{k=2}^{m+1}\binom{k+2n-c+i}{n-c+i+1} = \binom{m+2n-c+i+2}{n-c+i+2} - \binom{2n-c+i+2}{n-c+i+2}. \quad (14)$$

Application of this formula enables us to eliminate the inner sum in (11) so that

$$\sum_{k=2}^{m+1} A_{T_n,k} = \sum_{i=0}^{c-2} g_i(n)\left[\binom{m+2n-c+i+2}{n-c+i+2} - \binom{2n-c+i+2}{n-c+i+2}\right]. \quad (15)$$

The index k has been eliminated from the sum, and it only remains to show that the right-hand side of (15) is equivalent to $A_{T_{n+1},k}$ computed according to (7). This last step requires no summations, and after inserting the quantities obtained for $g_i(n)$, it is straightforward to complete algebraically.

3.2 Example: A Five-Taxon Case

We illustrate the approach using the T_5-caterpillar family in which the number of taxa in the left subtree of T_5 is 3 and the number in the right subtree is 2 (). For Step 1, the number of m-extended coalescent histories, as reported in [3, Table 1], is

$$A_{T_5,m} = \frac{1}{8}m(m^3+10m^2+31m+38). \quad (16)$$

For Step 2, by iterating (3), we obtain $A_{T_n,m}$ for $n=$ 6,7,8,9,10, as reported in Table 1.

The formulas for $A_{T_n,m}$ for $n=6,7,8,9,10$ are sufficient to propose the general pattern in Step 3 ($n_0=6$, $c=7$, and $n_0+c-3=10$). We see that each of these five formulas is a product of three components: a term $m/(n-1)!$, a term $(m+2n-7)!/(m+n-1)!$, and a polynomial of degree 4 in m. In this polynomial, we observe that the m^4 term has a constant coefficient of 3, the m^3 term is linearly increasing in n, the m^2 term is quadratically increasing in n, and so on. For each value of i from 0 to 4, $5-i$ values of n are required for identifying the unique polynomial of degree $4-i$ in n that passes through the coefficients of m^i for $n=6, 7, \ldots, 6+(4-i)$. For each of the terms m^i, we determine this polynomial in n, proposing a general formula for $A_{T_n,m}$:

$$A_{T_n,m} = \frac{m}{(n-1)!}\frac{(m+2n-7)!}{(m+n-1)!}[3m^4+(18n-48)m^3 + (41n^2-219n+283)m^2+(46n^3-369n^2 + 947n-774)m+(23n^4-246n^3+947n^2 - 1548n+896)]. \quad (17)$$

That $c=7$ is apparent from the formula.

For Step 4, to prove that the formula is correct, we must verify the proposed formula for $A_{T_n,m}$ by induction. The polynomial that plays the role of $\sum_{i=0}^{c-3}\sum_{j=0}^{c-3-i} a_{c-3-i,j}k^{i+1}n^j$ is $3k^5 + (18n-48)k^4 + (41n^2-219n+283)k^3 + (46n^3-369n^2+947n-774)k^2 + (23n^4-246n^3+947n^2-1548n+896)k$. Applying polynomial long division, this quantity can be rewritten as

$$\begin{aligned}&\frac{1}{(k+2n-7)!}[3(k+2n-2)! - 12(n-1)(k+2n-3)!\\&\quad + 17(n-1)(n-2)(k+2n-4)!\\&\quad - 8(n-1)(n-2)(n-3)(k+2n-5)!\\&\quad - 5(n-1)(n-2)(n-3)(n-4)(k+2n-6)!\\&\quad + 2(n-1)(n-2)(n-3)(n-4)(n-5)(k+2n-7)!].\end{aligned}$$

We can then write the proposed formula for $A_{T_n,k}$ as a sum

$$\begin{aligned}&3\binom{k+2n-2}{n-1} - 12\binom{k+2n-3}{n-2} + 17\binom{k+2n-4}{n-3}\\&\quad - 8\binom{k+2n-5}{n-4} - 5\binom{k+2n-6}{n-5} + 2\binom{k+2n-7}{n-6}.\end{aligned}$$

Using (14) to sum $A_{T_n,k}$ from $k=2$ to $m+1$, we have

$$\begin{aligned}\sum_{k=2}^{m+1} A_{T_n,k} &= 3\binom{m+2n}{n} - 3\binom{2n}{n}\\&\quad - 12\binom{m+2n-1}{n-1} + 12\binom{2n-1}{n-1}\\&\quad + 17\binom{m+2n-2}{n-2} - 17\binom{2n-2}{n-2}\\&\quad - 8\binom{m+2n-3}{n-3} + 8\binom{2n-3}{n-3}\\&\quad - 5\binom{m+2n-4}{n-4} + 5\binom{2n-4}{n-4}\\&\quad + 2\binom{m+2n-5}{n-5} - 2\binom{2n-5}{n-5}.\end{aligned}$$

It is then a matter of algebra to verify that this sum, equal to $A_{T_{n+1},m}$ by (3), is equal to the formula obtained from (17) by replacing n with $n+1$. Thus, the induction is complete, and (17) is established as the formula for the number of m-extended coalescent histories for the T_5-caterpillar-like family. Completing Steps 5 and 6 is then straightforward, producing

$$A_{T_n,1} = \frac{23n^2-131n+180}{4(2n-3)(2n-5)}C_{n-1}, \tag{18}$$

and

$$\lim_{n\to\infty}\frac{A_{T_n,1}}{C_{n-1}} = \frac{23}{16}. \tag{19}$$

3.3 All Choices of T_r with $r \le 8$ Taxa

For each T_r with $r \le 8$, I have employed the approach used in the five-taxon example together with the T_r-caterpillar-like family to propose a corresponding general formula for $A_{T_n,m}$. As computer algebra can often be employed to verify binomial identities [21] and the primary interest here is in comparing properties of the formulas for $A_{T_n,m}$ rather than in the proofs, I have verified by computer algebra rather than induction that $A_{T_{n+1},m} = \sum_{k=2}^{m+1} A_{T_n,k}$ for each choice of T_r with $r \le 8$.

For $r \le 7$, the formula determined in Step 3 for the number of m-extended coalescent histories appears in Table 2, and the number of coalescent histories and its limit appear in Table 3. As the formula for the number of m-extended coalescent histories is unwieldy for each T_r with $r = 8$ and our main interest is in $m = 1$, the formula for arbitrary m is omitted, and only the formula for the number of coalescent histories and the asymptotic limit are shown. For each T_r with $r = 8$, the number of coalescence sequences $N(T_8)$ for the labeled topology T_8 is also provided. In each case, $n_0 \le 12$.

Examining the limits $\lim_{n\to\infty}(A_{T_n,1}/C_{n-1})$, we can see that for each r from 4 to 8, the largest value occurs when the topology is r-maximally probable. For $r = 8$, Fig. 5 plots the limit as a function of the rank for topologies T_r. When T_r is not the minimal member of a caterpillar-like family, the limit for the appropriate minimal member with a lesser value of r is plotted. As each T_r with smaller r has some T_8 in its family, the plot can be viewed as illustrating the limits $\lim_{n\to\infty}(A_{T_n,1}/C_{n-1})$ for all T_r with $r \le 8$. For a given r, it is in this sense of extending families whose minimal member has size less than r to r that the largest value for the limit occurs when the topology is r-maximally probable.

Considering all 23 choices of T_r with $r = 8$, the limit is correlated with the rank, with correlation coefficient 0.846. It has local maxima at ranks 5 and 11, corresponding to the T_r-caterpillar-like trees in the families of the r-maximally probable topologies with $r = 6$ and $r = 7$. Additional local maxima occur at rank 9, where the left subtree has a five-maximally probable topology as one of its subtrees, and ranks 16 and 20, where the left subtrees are six- and five-maximally probable, respectively.

The limit $\lim_{n\to\infty}(A_{T_n,1}/C_{n-1})$ is even more strongly correlated with the number of coalescence sequences $N(T_8)$, with correlation coefficient 0.967. To transform $N(T_8)$, which ranges from 1 to 80, to a comparable scale to $\lim_{n\to\infty}(A_{T_n,1}/C_{n-1})$, which ranges from 1 to $\sim$5.39, Fig. 5 examines $1 + \ln N(T_8)$ for each T_8. This quantity and $\lim_{n\to\infty}(A_{T_n,1}/C_{n-1})$ are identical at the left endpoint of the plot (rank 1) and nearly identical at the right endpoint (rank 23), with $1 + \ln N(T_8)$ exceeding $\lim_{n\to\infty}(A_{T_n,1}/C_{n-1})$ by a mean of 0.834 across all ranks. The two quantities have the same pattern of increases and decreases, so that if $\lim_{n\to\infty}(A_{T_n,1}/C_{n-1})$ increases from rank i to rank $i+1$, then $N(T_8)$ also increases, and if $\lim_{n\to\infty}(A_{T_n,1}/C_{n-1})$ decreases from rank i to rank $i+1$, so does $N(T_8)$.

3.4 Corollaries

Rosenberg [3] obtained a lower bound on the ratio of the largest and smallest numbers of coalescent histories for matching gene tree and species tree labeled topologies with n leaves. That lower bound relied on the use of the n-taxon pseudocaterpillar as the tree with the largest known number of coalescent histories. As we have now established that for sufficiently large n, the n-taxon T_8-caterpillar-like tree using the eight-taxon symmetric tree as T_8 has a larger number of coalescent histories, we can replace the

TABLE 2
Number of m-Extended Coalescent Histories for T_r-Caterpillar-Like Families with $r \leq 7$

Number of taxa r	Topology T_r	General formula $A_{T_n,m}$ for the number of m-extended coalescent histories	Smallest n for which general formula holds (n_0)	Constant c in general formula
2		$\frac{m}{(n-1)!}\frac{(m+2n-3)!}{(m+n-1)!}$	2	3
4		$\frac{m}{(n-1)!}\frac{(m+2n-5)!}{(m+n-1)!}[2m^2+(5n-11)m+(5n^2-22n+21)]$	5	5
5		$\frac{m}{(n-1)!}\frac{(m+2n-7)!}{(m+n-1)!}[3m^4+(18n-48)m^3+(41n^2-219n+283)m^2+(46n^3-369n^2+947n-774)m+(23n^4-246n^3+947n^2-1548n+896)]$	6	7
6		$\frac{m}{(n-1)!}\frac{(m+2n-9)!}{(m+n-1)!}[4m^6+(39n-123)m^5+(157n^2-990n+1533)m^4+(338n^3-3198n^2+9883n-9963)m^3+(424n^4-5352n^3+24747n^2-49602n+36279)m^2+(306n^5-4830n^4+29728n^3-89028n^2+129338n-72570)m+(102n^6-1932n^5+14864n^4-59352n^3+129338n^2-145140n+65000)]$	8	9
		$\frac{m}{(n-1)!}\frac{(m+2n-9)!}{(m+n-1)!}[8m^6+(69n-237)m^5+(251n^2-1710n+2859)m^4+(499n^3-5064n^2+16784n-18099)m^3+(587n^4-7896n^3+38946n^2-83166n+64521)m^2+(405n^5-6780n^4+44324n^3-140904n^2+216595n-127752)m+(135n^6-2712n^5+22162n^4-93936n^3+216595n^2-255504n+119020)]$	8	9
		$\frac{m}{(n-1)!}\frac{(m+2n-7)!}{(m+n-1)!}[6m^4+(30n-90)m^3+(59n^2-357n+508)m^2+(58n^3-531n^2+1517n-1344)m+(29n^4-354n^3+1517n^2-2688n+1640)]$	6	7
7		$\frac{m}{(n-1)!}\frac{(m+2n-11)!}{(m+n-1)!}[5m^8+(68n-248)m^7+(401n^2-2923n+5252)m^6+(1342n^3-14667n^2+52639n-61994)m^5+(2804n^4-40856n^3+219675n^2-516173n+446795)m^4+(3800n^5-69220n^4+495610n^3-1741925n^2+3002057n-2026742)m^3+(3334n^6-72894n^5+651650n^4-3045850n^3+7840040n^2-10518408n+5732748)m^2+(1772n^7-45206n^6+484614n^5-2826850n^4+9675966n^3-19394772n^2+21019252n-9462696)m+(443n^8-12916n^7+161538n^6-1130740n^5+4837983n^4-12929848n^3+21019252n^2-18925392n+7181280)]$	10	11
		$\frac{m}{(n-1)!}\frac{(m+2n-11)!}{(m+n-1)!}[10m^8+(124n-484)m^7+(673n^2-5219n+10006)m^6+(2096n^3-24231n^2+92237n-115462)m^5+(4123n^4-63208n^3+358500n^2-890329n+815404)m^4+(5314n^5-101360n^4+761870n^3-2816725n^2+5112187n-3634126)m^3+(4478n^6-102102n^5+954295n^4-4672850n^3+12616699n^2-17758164n+10138884)m^2+(2320n^7-61558n^6+688032n^5-4192550n^4+15009024n^3-31469736n^2+35629436n-16700808)m+(580n^8-17588n^7+229344n^6-1677020n^5+7504512n^4-20979824n^3+35629436n^2-33401616n+13115376)]$	10	11
		$\frac{m}{(n-1)!}\frac{(m+2n-11)!}{(m+n-1)!}[15m^8+(180n-720)m^7+(942n^2-7506n+14754)m^6+(2826n^3-33630n^2+131508n-168744)m^5+(5362n^4-84706n^3+494180n^2-1259828n+1181727)m^4+(6682n^5-131390n^4+1016640n^3-3862300n^2+7187948n-5226840)m^3+(5466n^6-128478n^5+1236560n^4-6225800n^3+17250802n^2-24860866n+14494176)m^2+(2776n^7-75922n^6+873888n^5-5476400n^4+20125708n^3-43221558n^2+49987924n-23860176)m+(694n^8-21692n^7+291296n^6-2190560n^5+10062854n^4-28814372n^3+49987924n^2-47720352n+19009008)]$	10	11
		$\frac{m}{(n-1)!}\frac{(m+2n-9)!}{m+n-1)!}[10m^6+(84n-294)m^5+(294n^2-2058n+3514)m^4+(558n^3-5868n^2+19998n-22038)m^3+(624n^4-8772n^3+44802n^2-98322n+77908)m^2+(414n^5-7290n^4+49608n^3-162738n^2+256290n-153924)m+(138n^6-2916n^5+24804n^4-108492n^3+256290n^2-307848n+145224)]$	8	9
		$\frac{m}{(n-1)!}\frac{(m+2n-9)!}{(m+n-1)!}[20m^6+(150n-570)m^5+(480n^2-3630n+6650)m^4+(849n^3-9594n^2+34989n-40944)m^3+(897n^4-13476n^3+73431n^2-171066n+142694)m^2+(567n^5-10620n^4+76884n^3-267534n^2+444237n-278814)m+(189n^6-4248n^5+38442n^4-178356n^3+444237n^2-557628n+271764)]$	8	9

$(5n-12)/(4n-6)$ in the formula in [3, Theorem 3.18] with the corresponding term from the formula for the eight-taxon symmetric tree (see Table 3), or

$$\frac{1{,}381n^4-30{,}042n^3+244{,}979n^2-888{,}318n+1{,}209{,}600}{16(2n-3)(2n-5)(2n-7)(2n-9)}.$$

For large n, this result increases the upper bound in the ratio by a factor of $(1{,}381/256)/(5/4)=1{,}381/320\approx 4.32$.

Because $A_{S,1}=A_{S_L,2}A_{S_R,2}$ by (2), it is straightforward to calculate the number of coalescent histories for any tree S whose subtrees S_L and S_R both have a form for which the general formula for the number of m-extended coalescent

TABLE 3
Number of Coalescent Histories for T_r-Caterpillar-Like Families with $r \leq 8$

Number of taxa r	Topology T_r	Topology T_8 in the T_r-family	Rank of T_8	Number of coalescence sequences for T_8 ($N(T_8)$)	Number of coalescent histories $A_{T_n,1}$ for $n \geq n_0$	Asymptotic limit $\lim_{n\to\infty}(A_{T_n,1}/C_{n-1})$
2			1	1	C_{n-1}	1
4			2	2	$\frac{5n-12}{2(2n-3)}C_{n-1}$	5/4
5			3	3	$\frac{23n^2-131n+180}{4(2n-3)(2n-5)}C_{n-1}$	23/16
6			4	4	$\frac{3(17n^3-169n^2+542n-560)}{4(2n-3)(2n-5)(2n-7)}C_{n-1}$	51/32
			5	8	$\frac{3(45n^3-499n^2+1834n-2240)}{8(2n-3)(2n-5)(2n-7)}C_{n-1}$	135/64
			6	6	$\frac{29n^2-209n+360}{4(2n-3)(2n-5)}C_{n-1}$	29/16
7			7	5	$\frac{443n^4-6714n^3+37021n^2-87870n+75600}{16(2n-3)(2n-5)(2n-7)(2n-9)}C_{n-1}$	443/256
			8	10	$\frac{145n^4-2367n^3+14321n^2-38139n+37800}{4(2n-3)(2n-5)(2n-7)(2n-9)}C_{n-1}$	145/64
			9	15	$\frac{347n^4-5988n^3+38395n^2-108354n+113400}{8(2n-3)(2n-5)(2n-7)(2n-9)}C_{n-1}$	347/128
			10	10	$\frac{3(23n^3-279n^2+1096n-1400)}{4(2n-3)(2n-5)(2n-7)}C_{n-1}$	69/32
			11	20	$\frac{3(3n-16)(21n^2-171n+350)}{8(2n-3)(2n-5)(2n-7)}C_{n-1}$	189/64
8			12	6	$\frac{949n^5-20260n^4+168215n^3-677780n^2+1322676n-997920}{16(2n-3)(2n-5)(2n-7)(2n-9)(2n-11)}C_{n-1}$	949/512
			13	12	$\frac{2467n^5-55582n^4+493397n^3-2159642n^2+4667760n-3991680}{32(2n-3)(2n-5)(2n-7)(2n-9)(2n-11)}C_{n-1}$	2467/1024
			14	18	$\frac{3(489n^5-11486n^4+106701n^3-490076n^2+1112572n-997920)}{16(2n-3)(2n-5)(2n-7)(2n-9)(2n-11)}C_{n-1}$	1467/512
			15	24	$\frac{3355n^5-81574n^4+783749n^3-3715514n^2+8679984n-7983360}{32(2n-3)(2n-5)(2n-7)(2n-9)(2n-11)}C_{n-1}$	3355/1024
			16	48	$\frac{1157n^5-29723n^4+304213n^3-1551013n^2+3940266n-3991680}{8(2n-3)(2n-5)(2n-7)(2n-9)(2n-11)}C_{n-1}$	1157/256
			17	36	$\frac{3(327n^5-8251n^4+82716n^3-411736n^2+1017344n-997920)}{8(2n-3)(2n-5)(2n-7)(2n-9)(2n-11)}C_{n-1}$	981/256
			18	15	$\frac{5(127n^4-2286n^3+15065n^2-43146n+45360)}{16(2n-3)(2n-5)(2n-7)(2n-9)}C_{n-1}$	635/256
			19	30	$\frac{107n^3-1593n^2+7786n-12600}{4(2n-3)(2n-5)(2n-7)}C_{n-1}$	107/32
			20	45	$\frac{131n^4-2676n^3+20341n^2-68196n+85050}{2(2n-3)(2n-5)(2n-7)(2n-9)}C_{n-1}$	131/32
			21	20	$\frac{(n-4)(347n^3-5101n^2+24354n-37800)}{8(2n-3)(2n-5)(2n-7)(2n-9)}C_{n-1}$	347/128
			22	40	$\frac{(n-5)(487n^3-7411n^2+36954n-60480)}{8(2n-3)(2n-5)(2n-7)(2n-9)}C_{n-1}$	487/128
			23	80	$\frac{1381n^4-30042n^3+244979n^2-888318n+1209600}{16(2n-3)(2n-5)(2n-7)(2n-9)}C_{n-1}$	1381/256

histories is known. Rosenberg [3] examined bicaterpillars, shapes in which the subtrees on both sides of the root are caterpillars, finding that for an n-taxon bicaterpillar with caterpillar subtrees of sizes ℓ and $n-\ell$, $A_{S,1} = C_\ell C_{n-\ell}$. Generalizing this result, if for an n-taxon topology S, S_L is the ℓ-taxon topology in the caterpillar-like family starting at T_r and S_R is the $(n-\ell)$-taxon topology in the caterpillar-like family starting at T'_r, then by (2),

$$A_{S,1} = A_{T_\ell,2} A_{T'_{n-\ell},2} = A_{T_{\ell+1},1} A_{T'_{n-\ell+1},1}. \tag{20}$$

Thus, for any S, $A_{S,1}$ can be viewed in terms of the caterpillar-like families to which its left and right subtrees belong. Note that it immediately follows that if S is the minimal element in a (noncaterpillar) caterpillar-like-family—that is, if the two subtrees of the root of S each have two or more taxa—then $A_{S,1}$ is not a prime number, as it is a product of the numbers of coalescent histories for two trees of size 3 or more.

4 Discussion

We have evaluated the number of coalescent histories for cases in which the gene tree and species tree have the same labeled topology, belonging to a T_r-caterpillar-like family in which the subtree T_r has $r = 8$ or fewer leaves. For each T_r, we have obtained the result by following a general

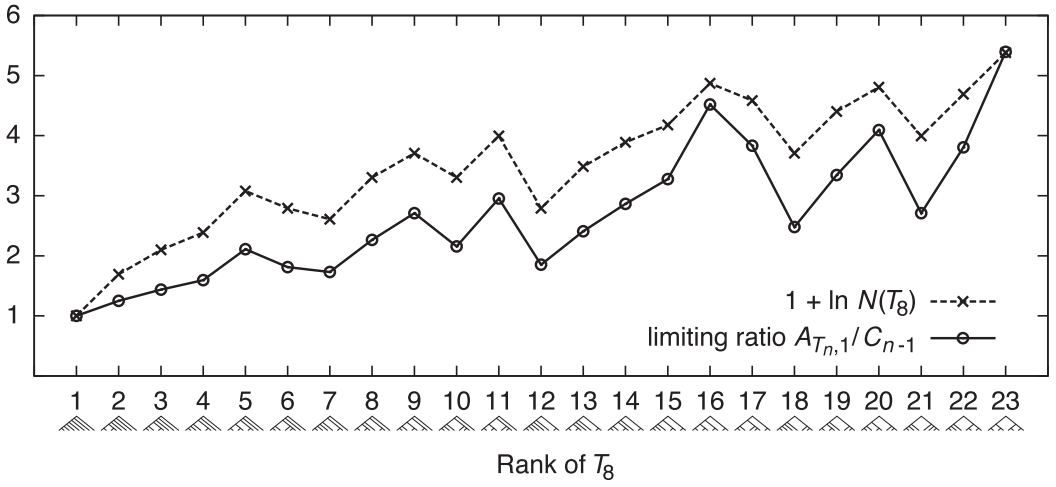

Fig. 5. The limiting ratio $\lim_{n\to\infty}(A_{T_r,1}/C_{n-1})$ for all T_r-caterpillar-like families with $r = 8$ leaves. The choices of T_r are listed with their ranks, and they follow the same order as in Table 3. A plot of 1 plus the natural logarithm of the number of coalescence sequences of eight-taxon labeled tree topologies (see (6)) is also shown.

procedure that extends a method previously used to obtain the number of coalescent histories in the caterpillar and pseudocaterpillar cases. This general procedure enables a formula to be proposed for the number of m-extended coalescent histories for T_r-caterpillar-like families, and then proven. While it has not been demonstrated that for any T_r, the procedure necessarily produces a general formula, in each of the small-r cases that have been examined, a formula that accurately reproduces values obtained according to the recursive equation (3) has been reported.

The numerical results support a previous claim that the number of coalescent histories for matching gene tree and species tree labeled topologies is large when the topology has both a large and relatively symmetric subtree and a high depth. The subtree generates a large number of sequences by which coalescences on the species tree can produce the gene tree, and the high tree depth provides a large number of branches on which those coalescences can occur. Among all T_8-caterpillar-like families, the largest number of coalescent histories occurs when T_8 is the fully symmetric eight-taxon subtree, approaching a limit of $1{,}381/256$ in relation to the number of coalescent histories for the n-taxon caterpillar. The results enable an answer to the question posed in Section 1, illustrating that the increased number of coalescence sequences when T_r is symmetric does compensate for the smaller depth of these trees: like the pseudocaterpillars, other T_r-caterpillar-like families have more coalescent histories than do caterpillars. The limiting number of coalescent histories at $r = 8$ generally increases with the rank of T_r, but an even more remarkable correlation is observed with the number of coalescence sequences $N(T_r)$. This relationship suggests that features of the number of coalescence sequences for a labeled topology, a quantity that has been studied for some time [16], [17], [18], [19], [20], [22], can provide an informal guide to properties of the limiting number of coalescent histories for caterpillar-like families.

It is important to clarify what has and what has not been shown. While each small T_r-caterpillar-like family studied led to a formula for the number of m-extended coalescent histories in the form given in (7), it has not been demonstrated that all T_r-caterpillar-like families have such a formula. It has, however, been shown, that if a formula can be proposed in the form presented in (7), then a general strategy exists for proving the formula by induction. This strategy of proof has been applied in full for one example case beyond the caterpillar and pseudocaterpillar cases that have been previously examined, and computer algebra has been used to verify the formulas in the remaining 20 cases.

The results provide a contribution to the study of gene trees and species trees, adding to the set of shapes for which detailed information is available about the number of coalescent histories. They will assist in relating the complexity of algorithms for computing gene tree probabilities based on coalescent histories to those that use a recursive evaluation based on a different class of objects, the "ancestral configurations" of Wu [23]. To more completely understand the properties of coalescent histories with n taxa, it will be of interest to extend beyond caterpillar-like families to obtain further results on the number of coalescent histories for trees of arbitrary shape.

Acknowledgments

The author thanks Lars Andersen for discussions of coalescent histories. This work was supported by US National Science Foundation (NSF) grant DBI-1146722 and by the Burroughs Wellcome Fund.

References

[1] J.H. Degnan and N.A. Rosenberg, "Gene Tree Discordance, Phylogenetic Inference and the Multispecies Coalescent," *Trends in Ecology and Evolution,* vol. 24, pp. 332-340, 2009.

[2] J.H. Degnan and L.A. Salter, "Gene Tree Distributions under the Coalescent Process," *Evolution,* vol. 59, pp. 24-37, 2005.

[3] N.A. Rosenberg, "Counting Coalescent Histories," *J. Computational Biology,* vol. 14, pp. 360-377, 2007.

[4] C. Than, D. Ruths, H. Innan, and L. Nakhleh, "Confounding Factors in HGT Detection: Statistical Error, Coalescent Effects, and Multiple Solutions," *J. Computational Biology,* vol. 14, pp. 517-535, 2007.

[5] N.A. Rosenberg and R. Tao, "Discordance of Species Trees with Their Most Likely Gene Trees: The Case of Five Taxa," *Systematic Biology,* vol. 57, pp. 131-140, 2008.
[6] E.S. Allman, J.H. Degnan, and J.A. Rhodes, "Identifying the Rooted Species Tree from the Distribution of Unrooted Gene Trees under the Coalescent," *J. Math. Biology,* vol. 62, pp. 833-862, 2011.
[7] C.V. Than and N.A. Rosenberg, "Consistency Properties of Species Tree Inference by Minimizing Deep Coalescences," *J. Computational Biology,* vol. 18, pp. 1-15, 2011.
[8] A. Hobolth, O.F. Christensen, T. Mailund, and M.H. Schierup, "Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model," *PLoS Genetics,* vol. 3, no. 2, article e7, 2007.
[9] A. Hobolth, J.Y. Dutheil, J. Hawks, M.H. Schierup, and T. Mailund, "Incomplete Lineage Sorting Patterns among Human, Chimpanzee, and Orangutan Suggest Recent Orangutan Speciation and Widepsread Selection," *Genome Research,* vol. 21, pp. 349-356, 2011.
[10] J.Y. Dutheil, G. Ganapathy, A. Hobolth, T. Mailund, M.K. Uyenoyama, and M.H. Schierup, "Ancestral Population Genomics: The Coalescent Hidden Markov Model Approach," *Genetics,* vol. 183, pp. 259-274, 2009.
[11] Y. Yu, C. Than, J.H. Degnan, and L. Nakhleh, "Coalescent Histories on Phylogenetic Networks and Detection of Hybridization Despite Incomplete Lineage Sorting," *Systematic Biology,* vol. 60, pp. 138-149, 2011.
[12] N.A. Rosenberg and J.H. Degnan, "Coalescent Histories for Discordant Gene Trees and Species Trees," *Theoretical Population Biology,* vol. 77, pp. 145-151, 2010.
[13] J.H. Degnan, "Gene Tree Distributions under the Coalescent Process," PhD dissertation, Univ. of New Mexico, 2005.
[14] J.H. Degnan and N.A. Rosenberg, "Discordance of Species Trees with Their Most Likely Gene Trees," *PLoS Genetics,* vol. 2, pp. 762-768, 2006.
[15] G.W. Furnas, "The Generation of Random, Binary Unordered Trees," *J. Classification,* vol. 1, pp. 187-233, 1984.
[16] J.K.M. Brown, "Probabilities of Evolutionary Trees," *Systematic Biology,* vol. 43, pp. 78-91, 1994.
[17] M. Steel and A. McKenzie, "Properties of Phylogenetic Trees Generated by Yule-Type Speciation Models," *Math. Biosciences,* vol. 170, pp. 91-112, 2001.
[18] E.F. Harding, "The Probabilities of Rooted Tree-Shapes Generated by Random Bifurcation," *Advances in Applied Probability,* vol. 3, pp. 44-77, 1971.
[19] J.M. Hammersley and G.R. Grimmett, "Maximal Solutions of the Generalized Subadditive Inequality," *Stochastic Geometry,* E.F. Harding and D.G. Kendall, eds., pp. 270-285, Wiley, 1974.
[20] E.F. Harding, "The Probabilities of the Shapes of Randomly Bifurcating Trees," *Stochastic Geometry,* E.F. Harding and D.G. Kendall, eds., pp. 259-269, Wiley, 1974.
[21] M. Petkovšek, H.S. Wilf, and D. Zeilberger, *A=B.* Peters, 1996.
[22] D. Aldous, "Probability Distributions on Cladograms," *Random Discrete Structures,* D. Aldous and R. Pemantle, eds., pp. 1-18, Springer-Verlag, 1996.
[23] Y. Wu, "Coalescent-Based Species Tree Inference from Gene Tree Topologies under Incomplete Lineage Sorting by Maximum Likelihood," *Evolution,* vol. 66, pp. 763-775, 2012.

Noah A. Rosenberg received the PhD degree in biological sciences from Stanford University in 2001 and completed his postdoctoral training at the University of Southern California. From 2005 to 2011, he was a faculty member at the University of Michigan. He is currently an associate professor in the Department of Biology, Stanford University. Research in his laboratory focuses on human evolutionary genetics, population-genetic theory, and mathematical phylogenetics.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.